



---

Supplement II to Recommendations for Reporting the Effectiveness of Programed Instruction Materials: Recommendations for Preparation of Technical Reports

Author(s): Joint Committee on Programmed Instruction and Teaching Machines

Source: *AV Communication Review*, Vol. 14, No. 2 (Summer, 1966), pp. 247-258

Published by: [Springer](#)

Stable URL: <http://www.jstor.org/stable/30217302>

Accessed: 21/06/2014 08:57

---

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at  
<http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



*Springer* is collaborating with JSTOR to digitize, preserve and extend access to *AV Communication Review*.

<http://www.jstor.org>

# Supplement II to Recommendations for Reporting the Effectiveness of Programed Instruction Materials

## *Recommendations for Preparation of Technical Reports*

(Revised October 1965)

JOINT COMMITTEE ON  
PROGRAMMED INSTRUCTION  
AND TEACHING MACHINES<sup>1</sup>

### I.

#### INTRODUCTION

##### A.

##### *Purpose and Scope*

This supplement contains further recommendations intended to serve as a guide for those who are preparing or reviewing technical documentation in support of statements about a program's performance characteristics—i.e., the outcomes that the program's use will demonstrably produce under specified conditions. Accordingly, these recommendations are concerned with what should be dealt with in technical reports of formal assessment studies in order that the general scientific criterion of reproducibility may be fulfilled.

For such studies, a detailed account of all experimental procedures and instruments used in assessment should be provided, giving the information needed for critical review of the study to determine whether summarized results and interpretations are warranted. The technical report on such detailed evaluation studies should permit a reviewer to assess the reported results in the light of the adequacy of criterion measures, description of test populations, description of conditions of experimental program use, and degree of correspondence

<sup>1</sup>Membership and sponsorship of the committee are indicated in footnote 2 of *Supplement I*.

between experimental conditions of testing and of those of intended application.

(Some of the considerations of experimental design and technical reporting dealt with here may not apply to informal program tryouts used only as guidance to the programmer in revising early versions of a program, nor to rough, preliminary screening tests made by a prospective user to help decide whether a published program seems to be generally suitable for his purposes.)

**B. Audience** This supplement is primarily addressed to behavioral scientists or educational research workers who provide technical assistance to the program user or producer in obtaining or interpreting comprehensive assessment data. This includes both those who prepare technical reports on effects of programs and those who advise purchasers regarding the soundness of reported data and statements concerning program effectiveness based on such data.

It is assumed that these individuals will have general competence in the fields of educational psychology, measurement, and experimental design. They should also be familiar with specialized technical considerations concerning program-assessment studies such as those discussed in papers by Lumsdaine<sup>2,3</sup> and by Jacobs, Maier, and Stolurow.<sup>4</sup>

**C. Main Considerations in Assessing Program Effects** Recommendations are given below for the four following elements in studies of program effects:

*1. Criterion Measures*

Behavioral indices of what students can do after going through a program, including definition of potential outcomes and their exemplification in appropriate criterion tests or other measures reflecting attainment of these outcomes.

*2. Characteristics of Students*

A complete description of the initial characteristics of the student population for which program-assessment data are being reported.

*3. Utilization Procedures and Experimental Design*

Procedures and arrangements for administering programs to defined samples of students under controlled and reproducible conditions, including procedures for administering the criterion tests.

*4. Processing, Analysis, and Reporting of Data*

Procedures for data processing and analysis, meeting scientific standards of reproducibility while also providing a basis for reporting the results in terms intelligible to the prospective program user.

<sup>2</sup> Lumsdaine, A. A. "Some Problems in Assessing Instructional Programs." *Prospectives in Programming*. (Edited by Robert Filep.) New York: The Macmillan Co., 1963. pp. 228-62.

<sup>3</sup> Lumsdaine, A. A. "Assessing the Effectiveness of Instructional Programs." *Teaching Machines and Programed Learning, II: Data and Directions*. (Edited by Robert Glaser.) Washington, D.C.: National Education Association, 1965. pp. 267-320.

<sup>4</sup> Jacobs, P. I.; Maier, M. H.; and Stolurow, L. M. *A Guide to Evaluating Self-Instructional Programs*. New York: Holt, Rinehart and Winston, 1966.

II.

**TECHNICAL  
RECOMMENDA-  
TIONS**

**A.  
Criterion  
Measures**

**1. Description of Tests**

The report should give a detailed description of content areas and corresponding behaviorally stated outcomes that were tested as possible effects of the program. These would include any instructional outcomes which were measured as possible effects of the programmer.

- a. *Standardized tests.* If standardized published tests are used, norms should be furnished, or published sources of normative data should be cited. If, in the context of reporting data in relation to such norms, data are also given for subsamples of items for which separate norms have not been published, the rationale and methods for selection of items should be explained, and the recommendations set forth below for nonstandard tests should be followed.
- b. *Nonstandardized tests.* Since standardized tests will generally not suffice to provide detailed evaluation of specific strengths and weaknesses of a program, specially constructed tests will commonly be used for program-evaluation studies. The considerations set forth below apply particularly to such nonstandard tests.
- c. Where appropriate, the use of relevant behavior samples (not limited to paper and pencil tests or other verbal measures) is encouraged. The procedures used in obtaining any such behavioral measures, of course, should be fully described.

**2. Detailed Identification of Test Content**

Copies of all test items used in measuring the possible outcomes that were tested should be included in the report if possible or, if not, should be available in an appendix. (Such appendix material as well as the basic report should be made permanently available from a suitable depository: University Microfilms, American Documentation Institute, etc., and the source of any such supplementary material should be cited in the technical report.) Test items may be presented in the form of complete specimen copies of the tests used in assessing the program effects. However, the test's content should also be described as accurately as possible in overall terms, and the test items, insofar as possible, should be keyed to the statements of specific categories of outcomes. Such an analysis should show which items from the test were used to measure each kind of outcome. For example, within the subject matter of algebra, one might identify the test questions used to measure the student's ability to solve quadratic equations of a given type.

**3. Scoring Keys**

The report (or appendix) should include not only copies of the test items themselves but also a specification of the answers considered to be acceptable and unacceptable for each. Otherwise the procedure will not be reproducible, and it will be difficult for a reviewer to ascertain just what a percentage of "correct answers" really means.

**4. Rationale for the Construction of Criterion Measures**

- a. *Definition of classes of outcomes.* The rationale for the construction or selection of test items and other measures of program effects developed for use in an assessment study should be reported as fully as possible. Such a rationale should include as comprehensive as possible a characterization of the entire class of behaviors which were represented or sampled in each particular test or subtest (including any classes of items used for the purpose of measuring "transfer"). Such definitions should be clarified by giving examples.
- b. *Sampling of items.* To the extent possible, the report should explain in what way the particular samples of test items employed in the study were generated. (In other words, the report should show the basis for determining the extent to which a person who does well on a particular set of test items would also be likely to do well on any other sample of items generated in a similar manner.) For example, in a program in spelling, the report should not only specify the particular words included in the test but also should describe the way in which the sample of test words was derived. As another example, various classes of quadratic equations might be identified, having specified ranges of coefficients and formats of expression with the samples of items used in the test(s) drawn from these according to a describable sampling plan.

It is recognized that for some kinds of subject matter, this kind of description must be quite imperfect at present because of limitations in the state of the art of behavioral taxonomy. However, the report should give as complete a description as is technically feasible at the present time so as to specify as accurately as possible the categories of outcomes that were tested.

5. *Independence of Samples of Items in Program and Test*  
 Many programs include only a relatively small sample of instructional items (frames) for a given objective; likewise, a feasible test of a program's effects will often contain only a relatively small sample of test items reflecting the kinds of outcomes to be assessed. In such cases, the report should indicate the procedures used to insure the independence of the sample of frames used in the program and the sample of items used in the criterion test (or tests). Specifically, it should show the extent of overlap and nonoverlap of specific examples used in the program and in the test. Also, it should state what precautions were taken to insure that the program did not merely coach the student on a particular sample of items used in the test to assess its effects. Although it is desirable for the universe of possible test items, or samples thereof, to be known to the programmer, the particular sample used in a criterion test should be unknown to him at the time the program is written if the program is designed to teach behaviors that are supposed to generalize to other similar items of behavior.

#### 6. *Measures of "Transfer"*

If the "transfer" value of the program to other types of performances is reported, the rationale and methods for measuring transfer should

be made explicit. As one example, an objective for a physics program might be to help students make new applications of principles not specifically dealt with in the program. Also, one might ascertain whether the program improved their ability to evaluate scientific experiments in other fields. The kind or degree of such "transfer" effects, if investigated, should be made explicit by description of the specific ways in which the "transfer" items differ from the content of the program.

(If the program's objectives and content are comprehensive enough to cover all relevant behavioral outcomes, all of its effects could be considered direct effects, so that the concept of transfer would not be applicable. However, this conception does not fit the case in which instructional frames comprise only a partial sample of the total class of relevant items of behavior.)

#### *7. Measures of Interest and Attitude*

Where data are presented on the extent to which student "interests" or "attitudes" are influenced by the use of a program, the report should present copies of the instruments used, including all specific questions asked to assess interests or attitudes. It should also specify the conditions of administering these instruments (including such methodological precautions as anonymity of responses).

Reports should make a clear distinction between data on students' interest in (or liking for) the program and gain in competence effected by the program. Their liking for the program itself should also be distinguished from interest aroused in the subject matter, and the student's liking for self-instructional programming generally should be distinguished from his liking of the particular program.

Reports should also reflect a distinction between behavioral indices of motivation or interest engendered by a program, such as students' volunteering to receive further instruction or being observed to engage in follow-up activities, versus mere verbal indicators believed to be predictive of such motivated behavior.

#### *8. Effects of Testing; Use of Parallel Test Forms*

The report should indicate the way in which the study took into account possible spurious effects of the testing procedure, including those resulting from use of the same test more than once for a given subject.

When parallel forms of measuring instruments are used (for example, the use of one form for a "before" test and one form for an "after" test, or the use of a parallel form in obtaining retention measures), their relationship to each other should be described fully. Any special techniques used (such as split forms with half of the group receiving one set of items before and the other half afterwards, while the reverse is true for another half of the group) should be explained in sufficient detail to be reproducible.

B.  
*Characteristics of Students*

1. *Comprehensive Description of Relevant Initial Competences*  
 The reader should be able to tell as precisely as possible what kinds of students were used in the study. They should be identified not only by

relevant general background or prerequisite characteristics but also in terms of their initial status with respect to the competences to be developed by the program.

*2. Detailed Data on Student Characteristics*

The report should identify in detail the characteristics of the students tested, including data on such factors as age, grade level, intelligence-test scores, reading ability, scholastic record, and initial competence of the kinds measured as outcomes. Other factors which it might be important to report, depending on the program, might include visual and auditory acuity and any required special aptitudes, such as manual dexterity. For such indices, appropriate measures of central tendency and spread (e.g., mean and standard deviation) should be supplied.

*3. Expected vs. Actual Student Competencies*

The report should indicate any substantial discrepancy between the expected prior level of competence (as indicated in advertising, program manual, etc.) and the extent of actual relevant prior competences possessed by the experimental subjects.

*4. Selection of Subjects*

The report should make clear how students were selected and assigned to the study. Reports should indicate how many students started and how many completed the program. For example, it should give relevant information about bases for potential selection bias, both in selection of schools or classes (e.g., a sample consisting only of those in which the teachers were willing to cooperate), and in individual self-selection of atypical students (e.g., the use of volunteers, bias due to dropouts). The characteristics of the dropouts should be reported in sufficient detail to determine the extent to which the remaining sample is representative.

The report should state explicitly what was done to deal with the problem of dropouts or absentees during the experiment. Measures for an attenuated subgroup should be accompanied by the earlier measures for that same subgroup so that the data for two or more time points are both based on a common sample present at both time points. However, the number and characteristics of absentees left out of the sample should be clearly identified. Their pretest score, when available, should be reported in relation to the scores of those who finished so as to reveal any biases that may have resulted.

*5. "Novelty Effects"*

The report should state the extent of students' prior experience with programmed materials (and/or presentation devices) of the kind whose effects are being reported.

C. The report should deal with two important aspects of "experimental design":

First, it should describe the *technical procedures and controls* employed so that the reader can assess the extent to which reported gains in knowledge, skills, etc., can be validly defended as results of the

program itself rather than of other concurrent or prior sources of influence.

Second, it should specify the *conditions of use of the program* which affect the *applicability* or generalizability of the results. For example, it should describe conditions of utilization in a classroom, or the use of a program for individual study, whether students were required to complete all of the frames or whether they merely had the program materials available to them to proceed with as far as they chose, etc.

**1. Generally Applicable Features**

To serve these purposes effectively, the report should deal with such details as the following:

- a. *The edition of the program used.* The extent to which the program used to collect data was different, if at all, from the commercially available edition. (If more than one edition is available, the report should specify which was used.)
- b. *Utilization situation.* The kind of situation under which the program was administered (for example, used in regular classrooms or in special settings). The conditions of the program's use should be reported in sufficient detail so that their essential features could be reproduced by another investigator.
- c. *Time intervals.* The distribution of amount of time per day students spent on the program, how long the program's use was continued, and time intervals between instruction and testing as well as between a first test and a later retention test. If not constant for all students, the distribution of such intervals should be given.
- d. *External help.* Any assistance supplied by the teacher or by others during the administration of the program, or at any time between the obtaining of pre- and post- (including retention) measures, should be fully reported. If teachers, or proctors, answered questions about the instructional content or procedures, full details should be given concerning control groups used to assess the effects of external assistance. (See also considerations applying to the use of control groups given below under "Comparative Studies.")
- e. *Motivational conditions.* The extent to which motivational influences were exercised, such as whether the teacher checked on students to make sure they were working on the programs or whether they were left to work by themselves, whether students were tested at intervals (and, if so, what tests were used and whether students were told that the tests counted on their grades), whether students were given any other special incentives for working on their programs or any disciplinary action for not working on them.
- f. *Testing conditions.* Conditions under which criterion tests were given, including: (1) total time taken for the tests (including *distribution* of times if variable) and whether the students were held to announced time limits; (2) instructions given to the students about the tests; (3) precautions taken to avoid inappropriate help

from teachers or other sources; and (4) any special incentives given in connection with testing.

- g. *Use of repeated measurements.* The report should indicate whether the same subjects were used for "before" measures as for immediate testing after the completion of the program and for retention data.
- h. *Need for preprogram measures.* If the experimenter has dispensed with a "before" measure (or equivalent measure from a separate uninstructed group) on the supposition that the student's initial level of knowledge is substantially zero, the basis on which this belief may be defended should be explicitly stated. In any case, the levels of attainment reported should in such instances be identified as measured competence following the program rather than as gain or effects due to the program.

#### 2. Comparative Studies

The following additional considerations apply to any studies in which data for two or more different treatment groups are compared.

- a. *Purpose of comparison.* The purpose of using comparison groups should be indicated—e.g., groups assigned to alternative programs or to alternative procedures for using a given program, or groups used as a control for extraneous sources of influence.
- b. *Definition of comparison treatments.* When the effectiveness of a program is being compared with the effectiveness of some other instructional procedure, full reporting of the nature of the "other" instruction, such as to make it substantially reproducible, is essential for valid interpretation.
- c. *Assignment of subjects to treatments.* The report should specify the procedures used to assign subjects to experimental treatments, e.g., by purely random assignment or random assignment of matched individuals.
- d. *Equivalence of groups.* Reports on studies in which equivalent sub-groups are used to obtain data at different time intervals should identify clearly the basis on which the comparability of these groups was established. Also, relevant activities intervening should be reported, and any interaction between the groups should be noted.
- e. *Control for confounded factors.* In any study in which the effects of one program or procedure are compared with those of another, procedural controls for insuring comparability of conditions for the two treatments should be reported in full, together with any known factors that might impair such comparability.

#### 3. Time vs. Criterion Achievement

A special problem in the assessment of self-instructional programs lies in the fact that there are two dependent variables of interest: (a) time spent in instruction and (b) proficiency, e.g., gain in achievement level. In the comparison of two programs, it is possible for one to produce higher achievement scores than the other, but also to require more time.

Gain in achievement level is sometimes expressed as an "efficiency" ratio of gain divided by time. Any such derived measure should be clearly explained so that such values as "percent efficiency" are not presented without the reader's being able to tell precisely what kind of derived measure was, in fact, employed. If no single achievement-time index seems defensible as a single figure of merit for a program's instructional efficiency, the alternatives for reporting are:

- a. Report gains in attainment of outcomes achieved or final levels of proficiency achieved by going through the program from beginning to end, separately reporting time spent on the program as a second dependent variable.
- b. Hold time constant experimentally, reporting attainment achieved in some arbitrarily fixed period of time, but preferably after two or three periods of time.
- c. Determine and report, as the main dependent variable, time required to achieve specified levels of attainment.

The third alternative presumes that all students reach some minimum level of proficiency. This involves repeated testing of each student's progress. Time-to-criterion can be employed as a sole dependent measure only if the basis for determining when the student has achieved criterion is based on such successive testing. Since the null hypothesis cannot be proved, it is *not* sufficient merely to show that two groups who took different amounts of time to complete alternative programs "did not differ significantly" with respect to the criterion level attained. Such statements should be scrupulously avoided.

**D.**  
*Analysis and Reporting of Data*

**1. General Considerations**

Results should be reported for all effects of the program which the study attempted to measure, including possible effects outside the primary objectives of the program and regardless of whether significant gains from the program were shown by the data.

**2. Analysis of Specific Program Effects**

Data for tests given before and after students have taken the program should be given for total scores and for content subscores so that a differential profile of program effectiveness can be made. For these measures, including time, the report should present summary statistics such as means and standard deviations. Such data should be given not only for the total group of subjects but also for subgroups differentiated by relevant student characteristics such as ability and initial knowledge. Such analyses should be accompanied by appropriate statistical tests of the reliability of any differential effects reported.

**3. Tests of Significance and Confidence Limits**

Enough information should be supplied to allow the reader of the technical report to check on the appropriateness of any inferential statistics reported—i.e., tests of significance or fiducial limits. Where the differences between two sets of scores are not statistically significant, the report should avoid the error of saying that the two sets of scores were

the "same" or "equivalent" results. Confidence limits for percentage values should be indicated when reporting for individual items and for means or other average score measures. The method by which confidence limits or significance tests are computed should be reported explicitly (since practice in computing such statistics is not uniform) so that an evaluator may verify the computation.

#### *4. Derived Measures*

The use of "percentage gain" or "percentage retention" measures, particularly when unqualified, is discouraged. Such measures should be accompanied by the basic data from which they are derived and by an explicit indication of how the percentage measure was obtained. They should also be accompanied by an indication of the standard error of such measures or by the data from which these standard errors can be derived.

#### *5. Reporting of Basic Data*

Any relevant details of assessment data which require more space than is appropriate for a published report should be made available (for example, in a supplementary report or by deposit with such agencies as the American Documentation Institute or University Microfilms). For example, supplementary tables should, whenever possible, be provided for the technical report, showing the complete matrix of all individual subjects' responses to all individual test items.

- a. Such an  $N \times k$  matrix ( $N = \text{no. of } Ss$ ,  $k = \text{no. of items}$ ) should be deposited with such an agency as ADI to permit checking and re-analysis of the data as desired by the technical reviewer.
- b. This information should be accompanied by each of the available scores of prerequisite knowledge and ability of each subject so that analysis can be made of results for ability subgroups other than those employed by the original report of the assessment study.

### III.

**CHECKLISTS** The checklists below are intended to recapitulate main points of the recommendations made in this supplement. In a well-reported assessment study, it should be possible to answer all of these questions in the affirmative. The information needed as a basis for such answers is indicated more fully by the recommendations given in Section II above. Affirmative answers to all questions do not guarantee the validity of the results of a study, but negative answers to any of the questions may call the validity of the study into question.

*Checklist "A":*

<i>Criterion Measures</i>	<ul style="list-style-type: none"> <li>(1) Does the report clearly identify the test instruments used to measure the behavioral effects of the program that were tested?</li> <li>(2) Are specimen copies of the tests and other measures supplied in the report or in an appendix?</li> <li>(3) Does the report supply and interpret the scoring key for all items?</li> <li>(4) Is the rationale for test content clearly specified both in terms of</li> </ul>
-------------------------------	---

behavioral categories and also by showing how the particular test items used to exemplify each category were generated?

- (5) Are there adequate safeguards against spurious effects due to selective coaching by the instructional program on specific items of the criterion tests?
- (6) Is the evidence clearly presented to indicate the nature of any effects reported for "transfer" to types of behavior not directly dealt with in the program?
- (7) Are instruments used to measure interest or attitudes provided, and are the conditions affecting their validity adequately described? Does the report clearly distinguish between effects on achievement and on interest or motivational effects?
- (8) Does the report deal adequately with special conditions affecting validity of measurement, including use of parallel test forms?

*Checklist "B":  
Characteristics  
of Subjects*

- (1) Has the report described clearly and completely the kind of student population with which the program was used which might influence the effectiveness of a particular program?
- (2) Similarly, has the report indicated what students were able to do, with respect to the outcomes tested, before they started the program?
- (3) Are the characteristics of test population and intended population substantially the same?
- (4) Has the report described how the schools and students were selected for the study so as to identify possible sources of selection bias, and does it deal adequately with such sources of selection bias, including bias due to dropouts?
- (5) Has the extent of students' prior experience with programs been taken into account?

*Checklist "C":  
Conditions of Use  
and Experimental  
Design*

- (1) *Conditions of reproducibility of program administration.* Does the report supply complete information regarding the way the program was used so that it could be administered again in the same way?
  - (a) Does the report indicate the form or edition of program that was used?
  - (b) Are the conditions under which the program was presented fully described?
  - (c) Are the time periods for program use and testing specified fully?
  - (d) Does the report describe fully the kind and amount of assistance supplied to students in the use of the programs?
  - (e) Are motivational conditions affecting students' work on the program adequately described?
  - (f) Does the report describe the conditions under which the tests were given, including use of repeated measures and any conditions which might alter the validity of the testing?
  - (g) Are numbers of cases tested fully reported, including identifi-

cation of any dropouts, and are measures for different time points based on equivalent samples of students?

- (2) *Validity of comparative studies*
  - (a) Where comparative results are given for alternative treatments, does the report describe the nature of the alternative treatments in a way that permits reproducibility?
  - (b) Does the report show that the students in the different comparison groups are equivalent samples, and does it describe the precise method of assigning students to alternative treatment groups?

*Checklist "D":  
Analysis and  
Reporting*

- (1) Does the report give completely and usefully the results of pre- and posttests and other evaluative measures?
  - (a) Are the results given not only for the total test but also for the subtests which indicate specific outcomes attained more or less effectively by the program?
  - (b) Are the test scores presented for each of the main subgroups in the student sample, and are the subgroups defined by relevant characteristics such as ability and background?
- (2) Are the methods for computing fiducial limits and tests of significance available so they can be verified?
- (3) Are all derived measures such as "percentage retention" and "percentage gain" clearly explained, and are the basic data on which they are based reported?